Stats 101C: Homework 2

Kitu Komya (404491375)

April 20, 2017

1.

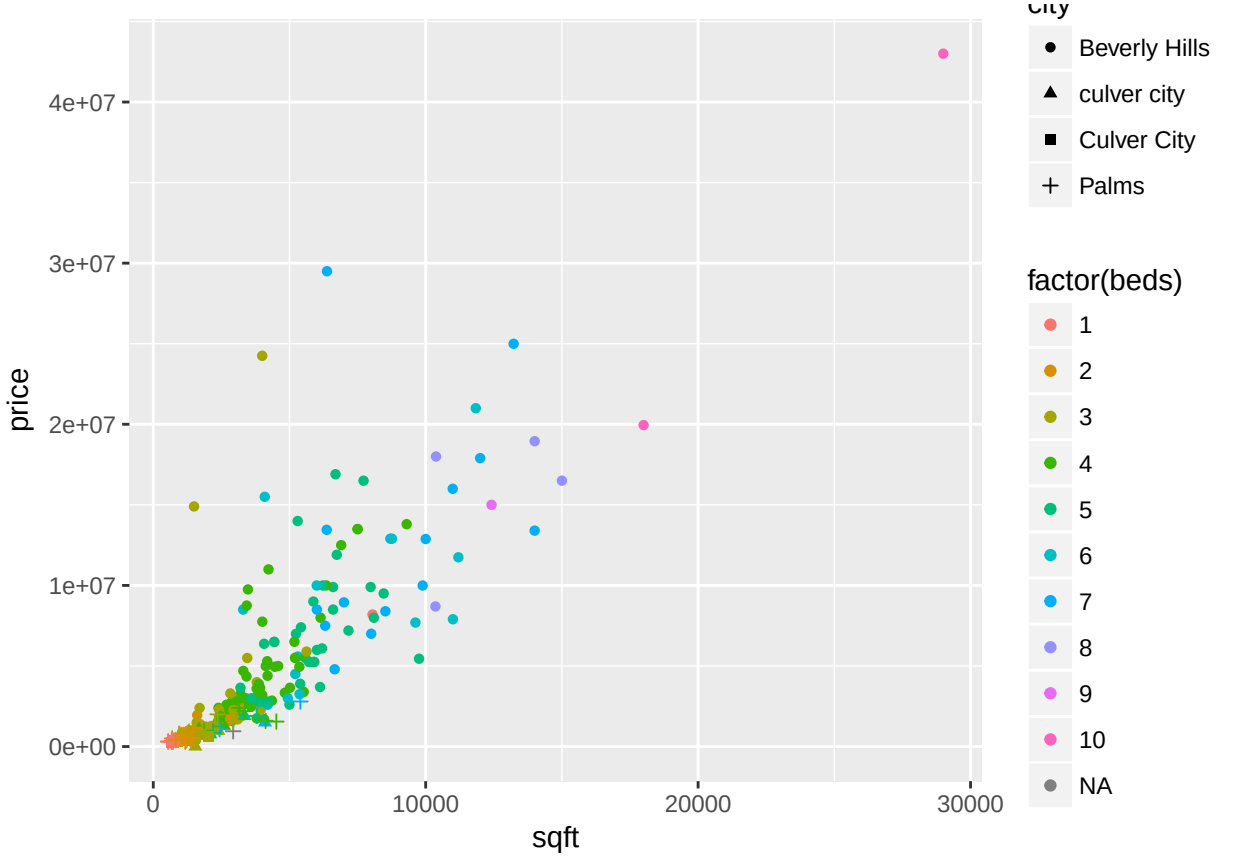
```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
lare <- read.csv(file = "LArealestate.csv")
ggplot(data = lare, aes(x = sqft, y = price, color = factor(beds), shape = city)) + geom_point()
```

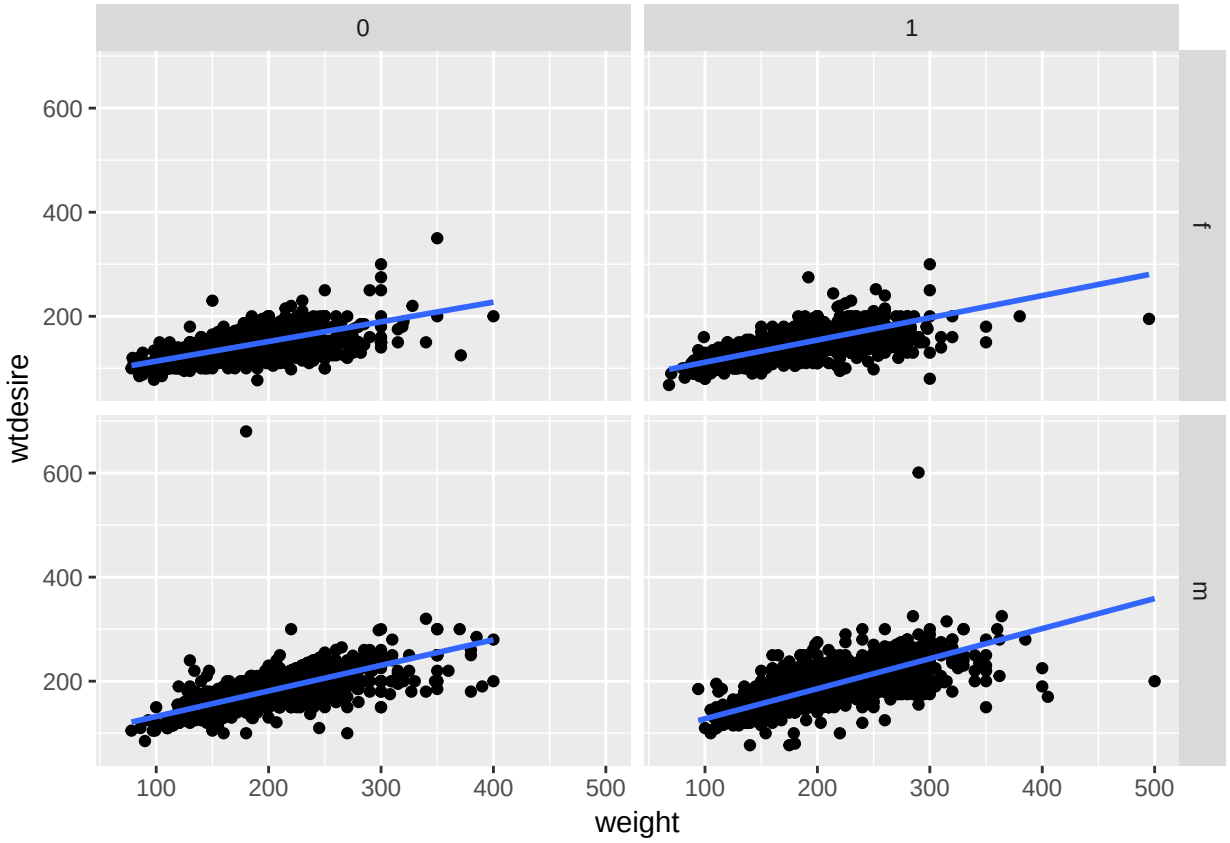
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Some questions this graph answers are: How does number of beds in a house relate to the price of the house? How does city relate to the price of the house? How does square footage relate to the price of the house?

2a.

```
cdc <- read.csv(file = "cdc.csv")
ggplot(cdc, aes(x = weight, y = wtdesire)) + facet_grid(gender~exerany) + geom_point() + geom_smooth(me
```

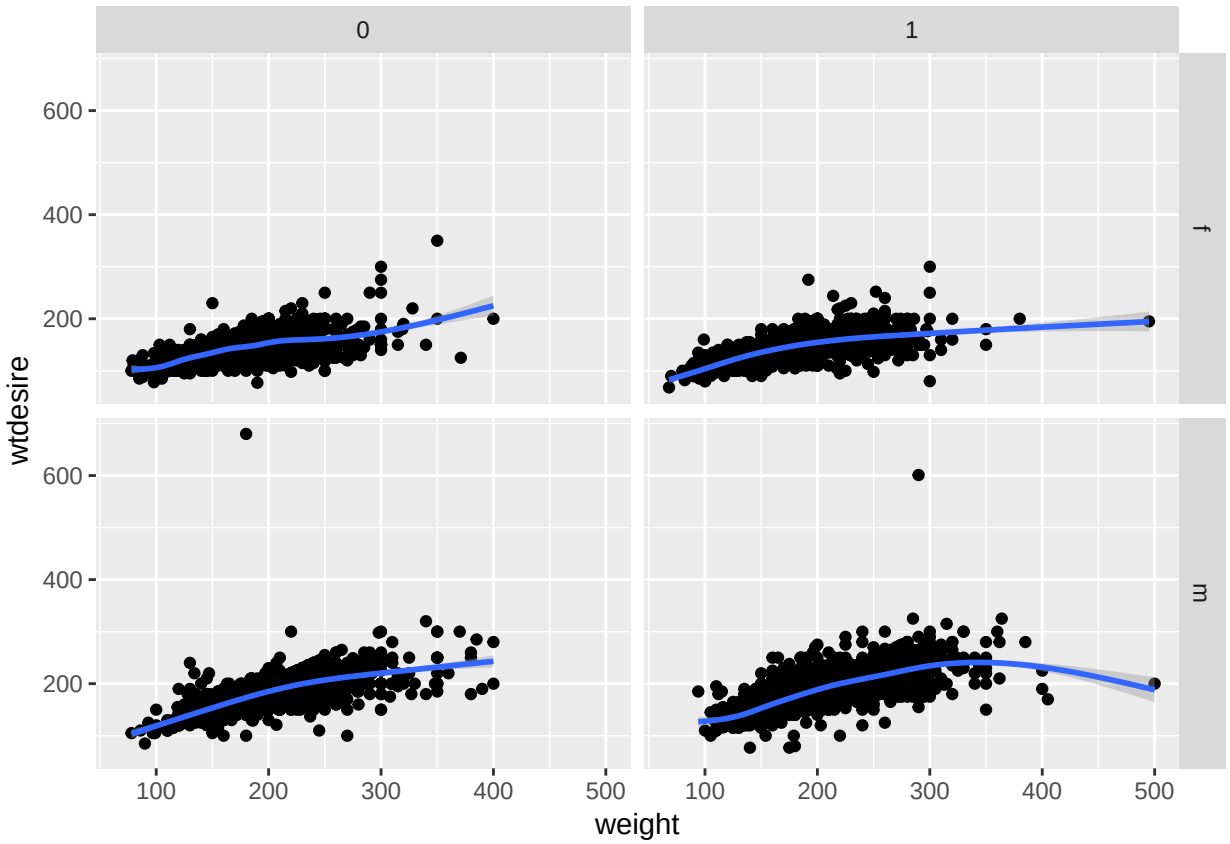


Among all 4 subgroups, we see that as weight increases, the desired weight also increases. The linear relationship is moderately strong and positive. Since all of the slopes of the four panels are approximately the same, it does not matter what gender you are or whether you exercise or not.

2b.

```
ggplot(cdc, aes(x = weight, y = wtdesire)) + facet_grid(gender~exerany) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```



The results differ now in that the lines are no longer quite as linear as they used to be. They are of some other polynomial type (a different flexibility). We still see that in general, as weight increases, the desired weight increases as well, but it no longer follows a linear relationship. In fact, for those who exercise and are male, it's almost a parabolic relationship, but we do note that it may have been corrupted by some outliers. In general however, we note that the flexibility of the model has changed (increased), while still retaining the same general relationship (positive association).

3.

banknote <- read.table(file = "banknote.txt", header = TRUE) normalize<-function(x) {return((x-min(x))/(max(x)-min(x)))} bn.norm <- cbind(as.data.frame(lapply(banknote[, 1:6], normalize)), factor(banknote$Y))

i = 1:dim(bn.norm) set.seed(33445566)

b.train <- sample(i, 140, replace = F) b.test <- sample(i, 60, replace = F)

bn.train <- bn.norm[b.train,] bn.test <- bn.norm[-b.train,]

library(class) model.1 <- knn(train = bn.train[, 1:6], test = bn.test[, 1:6], cl = bn.train[, 7], k = 1) model.1

model.3 <- knn(train = bn.train[, 1:6], test = bn.test[, 1:6], cl = bn.train[, 7], k = 3) model.3

model.5 <- knn(train = bn.train[, 1:6], test = bn.test[, 1:6], cl = bn.train[, 7], k = 5) model.5

library(caret) library(e1071)

trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3) set.seed(33445566) knn_fit <- train(banknote$Y ~., data = bn.train, method = "knn", trControl = trctrl, preProcess = c("center", "scale"),

tuneLength = 10)

tt <- data.frame(knn_fit[4]) plot(tt[,1],tt[,2],type="b",col="blue",xlab="k Nearest Neighbor",ylab="Accuracy") abline(v=3,col="red") qplot(tt[,1],tt[,2],xlab="k Nearest Neighbor",ylab="Accuracy",geom=c("point","smooth"),main="K vs Accuracy")

Accuracy was used to select the final model which was k = 3.

4a.

- Red: sqrt((0-0)^2 + (0-3)^2 + (0-0)^2)= sqrt(9) = 3
- Red: sqrt((0-2)^2 + (0-0)^2 + (0-0)^2)= sqrt(4) = 2
- Red: sqrt((0-0)^2 + (0-1)^2 + (0-3)^2)= sqrt(10) = 3.162278
- Green: sqrt((0-0)^2 + (0-1)^2 + (0-2)^2)= sqrt(5)= 2.236068
- Green: sqrt((0+1)^2 + (0-0)^2 + (0-1)^2)= sqrt(2) =1.414214
- Red: sqrt((0-1)^2 + (0-1)^2 + (0-1)^2)= sqrt(3)=1.732051

4b.

Green because the nearest single observation #5 is green.

4c.

Red because 2 of the 3 nearest neighbors to the test point are red.

4d.

Best value for K will be smaller because the boundary is non linear and will fit the data better.. When K becomes large, we get a smoother boundary.